

Examining the Effects of Social Stories on Challenging Behavior and Prosocial Skills in Young Children: A Systematic Review and Meta-Analysis

Charis L Wahman, PhD, BCBA-D
Assistant Professor
The Ohio State University
College of Education and Human Ecology
Department of Educational Studies
305 Annie and John Glenn Ave
Columbus, Ohio 43210-1172
(614)292-5468
price.1375@osu.edu

James E. Pustejovsky
Assistant Professor
The University of Texas at Austin
Educational Psychology Department
pusto@austin.utexas.edu

Michaelene M. Ostrosky, PhD
Professor
The University of Illinois at Urbana-Champaign
ostrosky@illinois.edu

Rosa Milagros Santos, PhD
Professor
The University of Illinois at Urbana-Champaign
rsantos@illinois.edu

August 2, 2019

Forthcoming in *Topics in Early Childhood Special Education*. This is the pre-print version. It is not the version of record and may not exactly replicate the final, published version of the article. The final article will be available, upon publication, via its DOI.

Abstract

Social stories are a commonly used intervention practice in early childhood special education. Recent systematic reviews have documented the evidence-base for social stories, but findings are mixed. We examined the efficacy of social stories for young children (i.e., 3-5 years) with challenging behavior across 12 single-case studies, that included 30 participants. The What Works Clearinghouse standards for single case research design were used to evaluate the rigor of studies that included social stories as a primary intervention. For studies meeting standards, we synthesized findings on the efficacy of social stories using meta-analysis techniques and a recently developed parametric effect size measure, the log response ratio. Trends in participants' response to treatment also were explored. Results indicate variability in rigor and efficacy for the use of social stories as an isolated intervention and in combination with other intervention approaches. Additional studies that investigate the efficacy of social stories as a primary intervention are warranted.

Keywords: challenging behavior, social stories, young children, intervention, meta-analysis, log response ratio

**Examining the Effects of Social Stories on Challenging Behavior and Prosocial Skills in
Young Children: A Systematic Review and Meta-Analysis**

Addressing children's challenging behavior has become a primary focus for practitioners, researchers, and policy makers (Hemmeter & Conroy, 2018; Ostrosky & Sandall, 2013). Limited social skills can result in challenging behavior which negatively impacts many areas of development, including children's self-confidence, relationships with peers and adults, self-regulation, ability to follow directions, and problem-solving skills (Hemmeter, Ostrosky, & Fox, 2006). Several factors are correlated with increased incidence of challenging behavior, including poor communication skills, delayed social and emotional skills, health issues, and environmental variables (Darling-Churchill & Lippman, 2016; Shonkoff, 2016). While approximately 10-15% of typically developing preschoolers exhibit mild to moderate levels of challenging behavior, this percentage is even greater among children from families living in poverty (Powell, Fixsen, Dunlap, Smith, & Fox, 2007). Young children exposed to multiple family risks factors are two to three times more likely to demonstrate aggression, anxiety and depression, and hyperactivity (National Center for Children in Poverty, 2009).

As the number of preschool children who live in poverty increases, more children will enter early childhood programs without the critical skills needed for successful school experiences. In fact, from 1995 to 2011 the percentage of children from low-income families enrolled in public or private preschools increased from 36 to 42% (Burgess, Chien, Morrissey, & Swenson, 2014). The failure to provide adequate social and emotional supports for children is not only costly for young children and their families, but for the community at large. In addition to the possibility of suspension and expulsion, preschoolers with challenging behavior often experience peer rejection and punitive interactions with adults, and they are at greater risk for

school failure (Gilliam, 2005; Hemmeter & Conroy, 2018). Alarmingly, challenging behaviors that appear early on in a child's life are predictive of adolescent delinquency, gang membership, and incarceration (Dodge, Bierman, Coie, Greenberg, Lochman, McMahon, et al., 2014; Huesmann, Dubow, & Boxer, 2009). Because of the developmental risks for children whose challenging behavior is not addressed early on (Losel & Bender, 2012; Tremblay, 2010), there is a need for interventions that support children who engage in challenging behavior.

Early childhood teachers need evidence-based intervention strategies as they work to address behavioral issues. The current review includes interventions that adhere to the specific prescribed story format by Gray and Garand (1993), as well as stories that describe expected behaviors and contexts. Social stories are one type of intervention that have been applied to decrease challenging behaviors and increase prosocial behaviors in young children (e.g., Benish & Bramlett, 2011; Lorimer et al., 2002; Rhodes, 2014). Several steps are required to create a Social Story™. It involves identifying a problematic social situation and target behavior, as well as establishing a context for the social situation. This information is gathered from observations of a target child and interviews with caregivers. Social Stories™ include six types of sentences: "a) descriptive, which identify the context of the target situation; b) directive, which describe a desired behavior in response to a social cue; c) perspective, which describe reactions or feelings in response to a social situation; d) affirmative, which express the value of a given context or culture; e) control, which provide analogies to promote understanding for the child; and f) cooperative, which include information about who will provide help and how that help will be made available for the child" (Sansosti, Powell-Smith, & Kincaid, 2004, p. 195). Gray and Garand (1993) recommend that a ratio of two to five descriptive, perspective, and/or affirmative sentences are used for every directive sentence in the story. The goal is to describe the social

situation and appropriate behaviors, rather than to direct the child about how to behave. Stories should be written at the child's comprehension level, clarity in print must be maintained, and vocabulary should be appropriate for the child (Gray & Garand, 1993). By including the aforementioned, children are able to grasp basic concepts and the social story is relevant to their needs.

To date, more than 15 reviews on the efficacy of social stories have been conducted (Garwood & Van Loan, 2017). Several reviews provided narrative syntheses on the effectiveness of this approach (e.g., Karkhaneh, Clark, Ospina, Seida, Smith, & Hartling, 2010; Rhodes, 2014); others applied a systematic analysis of the literature using both a methodological framework for evaluating quality, rigor, and effect size metrics for quantifying the effects of social story interventions (e.g., Karal & Wolfe, 2018; Leaf et al., 2015; Mayton, Menendez, Wheeler, Carter, & Chitiyo, 2013; McGill, Baker, & Busse, 2015; Qi, Barton, Collier, Lin, & Montoya, 2015; Zimmerman & Ledford 2017). Despite the use of quantitative analyses, previous systematic reviews have reached discrepant conclusions about the efficacy of social stories. For example, several researchers noted uncertainty about the efficacy of social stories for children with ASD due to weak treatment effects, confounding factors, inadequate participation, multi-component interventions, and poor study designs and implementation (Karkhaneh et al., 2010; Sansosti et al., 2004; Test, Richter, Knight, & Spooner, 2011). Others described social story interventions as "questionably effective" due to the variability of intervention outcomes observed based on Percentage of Non-Overlapping Data scores (Kokina & Kern, 2010). However, Karkhaneh et al. (2010) concluded that social stories were beneficial in modifying behaviors among high functioning children with ASD. Similarly, Rhodes (2014) found evidence for the effectiveness of social stories and Wong et al. (2014) identified this approach as an evidence-

based practice. It is important to note that previous systematic reviews have focused on children with Autism Spectrum Disorders (ASD), and not solely focused on young children with behavioral challenges. Qi et al. (2015) examined the effects of social stories using a variety of effect size measurements and determined that, although available evidence meets the criteria of the What Works Clearinghouse (WWC) 5-3-20 standard, social stories are not evidence-based according to their criteria for visual analysis. Using a quantitative analysis of effect sizes, Reynhout and Carter (2006) initially determined that social stories were a promising intervention, yet in a later synthesis they concluded that there is variation in the efficacy of social stories and that on average they are marginally effective.

Although some researchers have applied visual analysis procedures such as the WWC evidence standards (Kratochwill et al., 2013) and the Single Case Analysis and Review Framework (Ledford, Lane, Zimmerman, Chazin, & Ayres, 2016) to examine the quality and rigor of study designs, few studies have applied these methods in combination with parametric effect size measures and meta-analysis techniques. Compared to other techniques for summarizing evidence, parametric effect size measures and quantitative meta-analysis offer several potential benefits (Pustejovsky & Ferron, 2017). These statistical approaches provide a way to not only summarize findings about the magnitude of functional relations, but also to examine *variation* in treatment responses across participants and studies—allowing researchers to distinguish consistently effective interventions from ones that produce variable responses across individuals—and to identify characteristics that explain variation in treatment responses.

One challenge for applying meta-analysis methods to synthesize single-case studies is identifying suitable effect sizes for summarizing the magnitude of functional relations. Widely used indices such as the Percentage of Non-overlapping Data (Scruggs, Mastropieri, & Casto,

1987) and the Tau-U index (Parker, Vannest, Davis, & Sauber, 2011) have shortcomings that make them poorly suited for use in meta-analysis, including lack of comparability across studies that use different measurement procedures (Pustejovsky, 2018a; Tarlow, 2017) and unknown sampling distributions (Shadish, Rindskopf, & Hedges, 2008). In the present review, we applied a recently developed parametric effect size index called the log response ratio (Pustejovsky, 2015, 2018b), which has several advantages for synthesizing social story intervention studies. Specifically, the log response ratio is suitable for use with behavioral dependent variables measured through direct observation, which comprised the majority of outcomes in identified studies. Moreover, the log response ratio is closely related to the concept of percentage change from baseline, an intuitive and readily interpretable way to quantify the magnitude of functional relations.

In summary, the goal of this systematic review was to evaluate the efficacy of social stories to decrease challenging behavior and increase prosocial skills in children under age five by: a) assessing the quality of the available evidence using the WWC indicators; b) synthesizing findings using a parametric effect size and meta-analysis methods that are suitable for behavioral outcomes; and c) exploring potential moderators of treatment response.

Method

For this review, six online databases were searched (ERIC, Education Full Text, PsycArticles, PsychINFO, EBSCO, and CSA), and keywords included: *young children, preschool, Social Stories and scripted stories*. Following the online search, a hand search was conducted of the reference lists from key studies (Leaf et al., 2015; Qi et al., 2018; Wong et al., 2014; Zimmerman & Ledford, 2017). Studies had to meet the following criteria to be included in the review: (a) study was published in a peer-reviewed journal between 1995-2018; (b) study

reported one or more single-case designs (SCDs); (c) social stories were used as the primary intervention to reduce challenging behavior and increase prosocial behavior; (d) at least one intervention participant was under the age of 5 years; (e) child outcome data were presented for at least one measure of challenging behavior; and (f) study was conducted in the United States. Several studies meeting these criteria reported multiple SCDs (e.g., ABAB designs replicated across several participants). PRISMA guidelines were followed when conducting the literature search.

Coding Procedures

Intact SCDs identified for inclusion were coded based on descriptive characteristics, strength of research design, and strength of experimental control. We coded participant demographics (i.e., age, disability, gender, race/ethnicity) as reported in the articles. We also coded study design characteristics, including setting, type of single-case design, skills or behaviors targeted, presence of maintenance and generalization phases, and procedural fidelity defined as measurement in at least 33% of sessions and average scores higher than 80% across participants, conditions, and implementers (Barton, Meadan-Kaplansky, & Ledford, 2018).

The WWC standards (Kratochwill et al., 2013) for SCDs were used to assess 1) strength of research design (i.e., internal validity) and 2) strength of evidence for experimental control (i.e., visual analysis). Only intact SCDs that met WWC research design standards with or without reservations were used in the subsequent meta-analysis. Standards coded to evaluate the research design were: type of study design, systematic manipulation of the independent variable, repeated measurement of the dependent variable, inter-observer agreement (IOA) reported for more than 20% of data points in each condition, IOA higher than 80%, three attempts to demonstrate a treatment effect, at least three data points per phase, and an overall rating (e.g., meets WWC

standards, meets standards with reservations, does not meet standards). Standards coded “yes” or “no” to evaluate evidence of experimental control were: stability of baseline, overlapping data points, immediacy of change, consistency of change, evidence of a functional relation, and strength of the functional relation (e.g., no, moderate, or strong evidence).

The first author coded all designs (24 intact SCDs from 12 studies) that met the inclusion criteria. A doctoral student in special education was trained as a reliability coder on the aforementioned WWC standards in accordance with definitions provided by the WWC. Five of the twelve studies were randomly selected for reliability coding. Reliability was calculated by dividing the total number of agreements by the number of agreements plus disagreements and multiplying by 100. Average agreement calculated across all quality indicators and all studies was 93% (range = 74-100%). Both coders reviewed all disagreements and reached consensus.

In order to calculate effect sizes for the included designs, we used WebPlot Digitizer (Rohatgi, 2018) to extract outcome data from digitized versions of the single-case graphs presented in each article, a process that can yield highly reliable data (Moeyaert, Maggin, & Verkuilen, 2016). Extracted data were organized in an Excel spreadsheet.

Effect size calculations

The second author independently calculated parametric effect sizes for each case within each intact SCD, using data from phases that contrasted baseline conditions with a social stories intervention condition in its initial format. Data on the effects of modifications to the intervention were available only for a small sub-set of participants. As such, we excluded phases that involved modifications and provided a narrative review of the modifications. We also excluded one pair of phases from Burke et al. (2004) because the return-to-baseline phase consisted of a single data point.

For effect size calculations, we used the LRR-increasing form of the log response ratio (Pustejovsky, 2018b), so that positive values of the effect size correspond to improvements in behavior (i.e., reductions in disruptive behavior and improvements in pro-social behavior). For a behavior where improvement is desirable, the LRR-increasing is defined as $LRRi = \log(\mu_B/\mu_A)$, where μ_A is the average level of the behavior during baseline and μ_B is the average level of behavior during intervention. We used the bias-corrected estimator given in Pustejovsky (2018b) because some phases included only a few observations. Thus, we calculated LRR estimates for each dependent variable measured on each case, based on data from adjacent phases. In studies that had more than one pair of baseline and treatment phases, such as ABAB designs, we calculated LRR estimates for each pair of phases and then averaged the estimates using inverse variance weighting, resulting in one effect size estimate per case and dependent variable.

Meta-analysis

We conducted a meta-analysis of effect sizes from the cases in study designs that met WWC standards with or without reservations. To summarize the distribution of effect sizes across included cases and studies, we used a multi-level meta-analysis model (Pustejovsky, 2018b) that included random effects for studies and for cases nested within studies. We chose not to include random effects for intact designs because only a few studies included multiple intact designs. This model provides estimates of three key quantities: an overall average effect size, a case-level standard deviation (SD), and a study-level SD. The overall average effect size describes the average magnitude of behavior change due to intervention. However, if the effects of social stories vary from case to case, the average effect describes only part of the picture. Estimates of the case-level and study-level SD provide information to fill out that picture, by

describing the extent to which the effects vary from case to case (within a study) and across studies. Larger SD estimates indicate that effect sizes are more variable and less consistent.

We used cluster-robust variance estimation methods (CRVE; Hedges, Tipton, & Johnson, 2010) with small-sample adjustments (Tipton & Pustejovsky, 2015) to calculate standard errors and confidence intervals for overall average effect size estimates, with clustering at the study level. This method is robust to mis-estimation of the standard errors of individual LRR estimates, as could occur if there is an auto-correlation or trend in the data series. We report several aids for interpreting the meta-analysis results. First, we translate overall average LRR effect sizes into percentage change terms, using the relationship $\% \text{ Change} = 100\% \times [\exp(S \times LRR) - 1]$, where S is equal to 1 for desirable behaviors and -1 for undesirable behaviors and $\exp(\cdot)$ denotes the natural exponent function (Pustejovsky, 2018b). Second, to interpret the magnitude of the within- and between-study SD estimates, we report 67% prediction intervals (PI) for individual effect sizes in percentage change terms. Prediction intervals have been recommended as a clinically interpretable description of effect size distributions (Borenstein, Higgins, Hedges, & Rothstein, 2017). The 67% PI characterizes the range of responses that we would anticipate for 2/3 of the population, if one were to use the social stories intervention with new participants. Wider PIs indicate more heterogeneous responses to treatment.

Finally, in addition to estimates of overall average effect sizes, we conducted meta-regression analyses to explore whether participant characteristics or study design features explain variation in the magnitude of effect sizes. We examined four potential moderators: participant age, participant diagnosis, whether the interventionist was also the primary data collector, and overall WWC design rating. We report separate meta-regressions for each moderator, pooling

across challenging behavior and pro-social behavior due to the small number of studies that include dependent variables of each type. For purposes of examining WWC design ratings as a moderator, we included studies that did not meet WWC standards.

All of the analyses were conducted in the R statistical computing environment (Version 3.5.1; R Core Team, 2018), using the SingleCaseES package for effect size calculations (Pustejovsky & Swan, 2018), the metafor package for meta-analysis (Viechtbauer, 2010), and the clubSandwich package for robust variance estimation (Pustejovsky, 2017). Raw data and R scripts for replicating the analyses are available at <https://osf.io/7uvr3/>.

Results

Participants. Database searches led to the identification of 257 studies from 1995-2018. Two hundred and forty-three studies were excluded following title and abstract review. Two additional studies were identified by searching the references from other studies and literature reviews. Twelve studies were examined at the full text level and met the inclusion criteria. The 12 identified studies included 24 intact SCDs and 30 participants. Figure 1 displays a PRISMA diagram (Moher, Liberati, Tetzlaff, & Altman, 2009) that summarizes the screening process.

Table 1 provides a summary of the 12 studies that met all inclusion criteria and were assessed using WWC standards. Across the studies, participating children ranged in age from 2:6 to 10:3 years with a mean age of 5:3 years. Twenty-five children were between 3 and 5 years old. Two children were female and 23 were male. Of the participating children, 22 children were identified as having special needs, which included ASD, Developmental Delay (DD), and Specific Language Impairment (SLI). Seven children were not identified as having special needs. Specifically, two studies included typically developing children (Benish & Bramlett, 2011; Burke, Kuhn, & Peterson, 2004), and one study included a participant who exhibited hyperlexia,

an advanced reading ability (Soenksen & Alper, 2006). Six research teams reported the ethnicity of their participants, which included Hispanic, Chinese, Caucasian, and African American (Burke et al., 2004; Chan & O'Reilly, 2008; Hsu, Hammond, & Ingalls, 2012; Ivey, Heflin, & Alberto, 2004; Kuoch & Mirenda, 2003; Schneider & Goldstein, 2010).

Settings. Included studies were conducted in a variety of settings. Seven research teams conducted social story interventions in classrooms (Benish & Bramlett, 2011; Chan & O'Reilly, 2008; Crozier & Tincani, 2007; Hsu et al., 2012; Schneider & Goldstein, 2010; Soenksen & Alper, 2006; Wright & McCathren, 2012). One study was conducted in both home and classroom environments (Kuoch & Mirenda, 2003), while another was conducted in both the participants' home environment and a university research room (Leaf, Oppenheim-Leaf, Call, Sheldon, & Sherman, 2012). Two studies were conducted only in home environments (Burke et al., 2004; Lorimer et al., 2002); one study was conducted in a clinic (Ivey et al., 2004).

Target skills. Across all studies, the goal was to decrease challenging behaviors (e.g., avoidance, physical aggression, name calling, tantrums, destruction of property, crying, yelling, making negative comments, disruptive bedtime behaviors), and/or increase prosocial behaviors (e.g., raising hand, saying a peer's name, looking at peer's face, sitting appropriately during circle time, and following directions).

Maintenance and generalization. Only one research team reported both generalization and maintenance data (Leaf et al., 2012). Seven studies reported maintenance data.

Multi-component interventions. Three studies examined the implementation of social stories as a packaged intervention (Burke et al., 2004; Chan & O'Reilly, 2008; Crozier & Tincani, 2007). For these studies, additional components included verbal prompts, role-play, and positive rewards. Leaf et al. (2012) compared social stories to a teaching interaction procedure

and data on these two interventions were reported in separate single case design graphs. For the purpose of this review, only data on the social story intervention were analyzed.

Implementation of Procedural Fidelity. Although 10 research teams discussed implementation of procedural fidelity procedures, studies varied in reported procedures and outcomes. Two studies did not report procedural fidelity data (Lorimer et al., 2002; Soenksen & Alper, 2006). Across the 10 studies that reported outcomes, the average procedural fidelity was 99% (range = 96-100%).

Strength of Research Design

Study Design. Eight studies included multiple baselines across participants or behaviors, while six included at least one reversal design. Seven studies contained multiple designs. Table 1 reports overall assessments of strength of research design and evidence of experimental control for each design. Tables S1 and S2 in the accompanying supplementary materials report ratings for the specific criteria that inform these assessments.

Independent and Dependent Variables. Researchers implemented the independent variable, or intervention, before participants entered the target setting where behavior was observed. For example, an interventionist read the social story to the participants prior to them entering the target setting where challenging behavior typically occurred (Lorimer et al., 2002; Soenksen & Alper, 2006). Researchers also selected a variety of measurement systems to assess dependent variables. For instance, Lorimer et al. (2002) and Soenksen and Alper (2006) used event recording to measure the frequency of occurrence of target behaviors during baseline and intervention phases.

Inter-observer Agreement. IOA data were reported for more than 20% of the data points in each condition at more than 80% for 11 of the 12 studies. Across the 11 studies, the

average IOA was 94.24% (range = 81-100%). In one study the authors indicated that researchers were trained on IOA, however IOA data were not reported (Benish & Bramlett, 2011).

Potential Demonstrations of Effect. Twenty-three intact designs from 11 studies provided three attempts to assess potential treatment effects.

Data Points Per Phase. All 12 studies included at least one design that had three or more data points per phase, although two studies included at least one design that had less than three data points per phase (Burke et al., 2004; Crozier & Tincani, 2007). Neither study highlighted the number of data points per phase as a limitation. Only four of the 12 studies included one or more designs that had at least five data points per phase (Crozier & Tincani, 2007; Kuoch & Mirenda, 2003; Schneider & Goldstein, 2010; Wright & McCathren, 2012).

Overall rating. Designs were scored as not meeting minimal standards if they did not adhere to the minimal criteria described earlier. Three studies contained at least one design that met the WWC standards (Crozier & Tincani, 2007; Schneider & Goldstein, 2010; Wright & McCathren, 2012), while seven studies contained at least one design that met the standards with reservations. In contrast, five studies contained at least one design that did not meet the WWC standards.

Evidence of Experimental Control

Each study design was evaluated to determine if a relation between the independent variable and an outcome variable was demonstrated. Each design was scored as: 1) *no evidence* if it did not provide at least three demonstrations of an effect; 2) *moderate evidence* if at least one demonstration of a non-effect was evident; or 3) *strong evidence* if it provided three or more demonstrations of effect and no evidence of non-effects.

Stable baseline. Four studies contained at least one design that demonstrated a stable baseline (Crozier & Tincani, 2007; Leaf et al., 2012; Soenksen & Alper, 2006; Wright & McCathren, 2012), with three data points demonstrating minimal variability and a consistent level throughout baseline. Eight studies demonstrated unstable baselines due to variability and inconsistency in the trend of the data points within each phase and across phases. For example, in Chan and O'Reilly (2008), baseline data were highly variable. Additionally, prior to introducing the intervention phase, a downward trend in baseline data was evident.

Overlapping Data Points. All 12 studies contained overlapping data points across at least one phase of the design. In Lorimer et al. (2002), two baseline data points overlapped with two intervention data points. When the baseline phase was repeated, an overlap in data points was still evident. This pattern also was observed in the Crozier and Tincani (2007) study. In fact, for their second participant, data points from baseline to the first intervention phase overlapped significantly, which resulted in no changes in behavior across those phases for this participant.

Immediacy and Consistency of Change. To determine immediacy of an effect, a change in level between the last three data points of one phase and the first three data points of the next phase should be evident. Seven studies contained at least one design that demonstrated an immediate effect of intervention on target behaviors (Benish & Bramlett, 2011; Burke et al., 2004; Chan & O'Reilly, 2008; Crozier & Tincani, 2007; Kuoch & Mirenda; 2003; Lorimer et al., 2002; Schneider & Goldstein, 2010). For example, Chan and O'Reilly (2008) demonstrated an immediate change in level between the last three baseline data points and the first three intervention data points. However, this change in level was consistent only for the first two tiers of the design. Crozier and Tincani (2007) demonstrated similar results for two participants. An immediate and consistent effect was evident for a third participant following the implementation

of a second intervention, verbal prompts. Three of the seven studies failed to demonstrate consistency of change across phases and conditions (Benish & Bramlett, 2011; Kuoch & Mirenda, 2003; Lorimer et al., 2002).

Evidence of Functional Relation. Six studies contained at least one design that demonstrated a functional relation between the independent and dependent variables (Benish & Bramlett, 2011; Burke et al., 2004; Crozier & Tincani, 2007; Lorimer et al., 2002; Schneider & Goldstein; 2010; Soenksen & Alper, 2006), while 10 studies did not demonstrate a functional relation for at least one of their designs (Benish & Bramlett, 2011; Burke et al., 2004; Chan & O'Reilly, 2008; Crozier & Tincani, 2007; Hsu et al., 2012; Ivey et al., 2004; Kuoch & Mirenda, 2003; Leaf et al., 2012; Lorimer et al., 2002; Wright & McCathren, 2012); these studies did not meet this criterion because of unstable baselines, overlapping data points across baseline and intervention, and no intervention phase. Additionally, an immediate change in the dependent variable was not evident when the independent variable was introduced or changes in the dependent variable were not consistent across repeated phases (i.e., baseline, intervention).

Strength of Relation. Six studies contained at least one design that demonstrated moderate evidence of experimental control (Benish & Bramlett, 2011; Burke et al., 2004; Crozier & Tincani, 2007; Lorimer et al., 2002; Schneider & Goldstein, 2010; Soenksen & Alper, 2006), while nine studies contained at least one design that did not demonstrate experimental control (Benish & Bramlett, 2011; Burke et al., 2004; Chan & O'Reilly, 2008; Crozier & Tincani, 2007; Hsu et al., 2012; Ivey et al., 2004; Kuoch & Mirenda, 2003; Leaf et al., 2012; Wright & McCathren, 2012). All cases demonstrated moderate evidence for three studies (Lorimer et al., 2002; Schneider & Goldstein, 2010; Soenksen & Alper, 2006) and no evidence for six studies (Chan & O'Reilly, 2008; Hsu et al., 2012; Ivey et al., 2004; Kuoch & Mirenda, 2003; Leaf et al.,

2012; Wright & McCathren, 2012). No studies demonstrated strong evidence of experimental control.

Meta-analysis and meta-regression

Table 2 reports results of the overall meta-analysis of LRRi effect sizes, including estimates of overall average effect sizes, study-level variation, and case-level variation, based on all included studies and cases. For challenging behavior, the average LRRi estimate was 1.22, which corresponds to a reduction of 70% from baseline levels, 95% CI [-93%, 22%] and was not statistically distinguishable from a null average effect. For pro-social behavior, the average LRRi estimate of 0.94 corresponds to a 155% improvement, 95% CI [56%, 318%], which was statistically distinguishable from null. However, there was substantial variation in the effects across cases for both types of outcomes, as well as across studies for challenging behavior outcomes. Accounting for case- and study-level variation, a 67% PI for challenging behavior ranges from a 15% to 90% reduction. A 67% PI for pro-social behavior ranges from an 16% *reduction* (i.e., iatrogenic effect) to a 677% improvement from baseline.

We conducted separate meta-regression analysis of four potential moderators, including WWC study design rating, participant age, participant diagnosis, and whether the interventionist was also the primary data collector (see Table 2). None of the four moderators explained a statistically significant degree of variation in the effect size estimates. Although the differences are not statistically distinguishable, it is worth noting that average effect size estimates were smaller (i.e., less beneficial effects) for 1) participants with diagnosed disabilities, 2) studies that met WWC design standards without reservations, and 3) studies where the interventionist was not also the primary data collector. After controlling for all four moderators using a join meta-regression, the average LRRi effect sizes for challenging behavior and prosocial behavior were

reduced and imprecisely estimated. See Table S3 of the supplementary materials.

Multi-Component Interventions

Similar to studies that implemented social stories in isolation, data on the four studies that implemented social stories as a package were variable in terms of rigor and effectiveness. While all four studies demonstrated low to medium rigor for at least one of their designs (Burke et al., 2004; Chan & O'Reilly, 2008; Crozier & Tincani, 2007; Schneider & Goldstein, 2010), only two demonstrated high rigor (Crozier & Tincani, 2007; Schneider & Goldstein, 2010).

Discussion

The efficacy of social stories in decreasing challenging behavior and increasing prosocial skills was evaluated by assessing the quality of available evidence using the WWC indicators and synthesizing the findings using a parametric effect size and meta-analysis methods. We also explored trends in participant response to treatment based on participant diagnoses, participant age, WWC design rating, and primary data collector. Overall, results indicate variability in rigor and effectiveness for the use of social stories as an isolated intervention and in combination with other intervention approaches.

For the studies that met the minimal WWC standards, we found that social story interventions for preschoolers had variable effects on challenging behavior and prosocial behavior. Several studies contained at least one design that did not meet minimal WWC standards and did not demonstrate experimental control. It is important to note that although studies may provide evidence of a strong design, they may still fail to demonstrate evidence of a causal relation between independent and dependent variables. Only six studies contained at least one design that met the standards or met them with reservations while also demonstrating experimental control. It is important to interpret these findings with caution, as data were highly

variable for five of these six studies (Chan & O'Reilly, 2008; Crozier & Tincani, 2007; Lorimer et al., 2002; Schneider & Goldstein, 2010; Soenksen & Alper, 2006). Hence, the effectiveness of social story interventions is uncertain because several studies did not adhere to the WWC standards, and variability in the data made it difficult to reach a definitive conclusion based on visual analysis.

Parametric effect size calculations estimated an average reduction in challenging behavior and improvement in prosocial skills for participants from baseline levels, although substantial variation was observed across cases for both outcomes. Surprisingly, little of this variation was explained by the four moderator variables examined (i.e., participant diagnosis, participant age, WWC design rating, primary data collector). It is also notable that average effect size estimates were smaller in magnitude when controlling for these moderators, although none of the individual moderators were statistically significant. That is, studies conducted with high rigor tended to exhibit lower effect size magnitude than studies with low rigor. Additionally, average effect size estimates tended to be smaller for participants with disabilities and studies conducted where the interventionist was not also the primary data collector. These trends are worrisome because they suggest potential biases that might impact the integrity of study outcomes. Thus, additional studies conducted with high rigor are needed to get reliable estimates of effect sizes and to understand factors that explain variation in treatment response.

A concern worth highlighting is the presence of multiple-component interventions without an attempt to evaluate the efficacy of social stories in isolation. While eight research teams implemented social stories as a singular intervention, four teams combined social stories with other interventions or instructional methods, such as verbal prompts, rewards, visual schedules, or role play (Burke et al., 2004; Chan & O'Reilly, 2008; Crozier & Tincani, 2007;

Schneider & Goldstein, 2010). Two multi-component designs met the WWC standards with reservations and provided moderate evidence for experimental control (Burke et al., 2004; Chan & O'Reilly, 2008). Although two additional multi-component designs did not meet minimal standards due to the strength of their design, they demonstrated moderate evidence of experimental control (Crozier & Tincani, 2007; Schneider & Goldstein, 2010). As such, multi-component interventions warrant further investigation to consider the additive effect of intervention components on participant behavior. Researchers should attend to the contributions of each intervention component to determine whether behavioral changes are the result of social stories, the impact of other interventions, or a combination of both.

Limitations. This review is one step towards evaluating the efficacy of social stories for young children who engage in challenging behaviors. One limitation of our review is that we focused only on studies in which challenging behavior and prosocial skills were the target behaviors. It is important to examine the effectiveness of social stories when implemented for other behaviors such as adaptive skills and oral communication (Kassardjian et al., 2014; Laprime, & Dittrich, 2014; Raver, Bobzien, Richels, Hester, & Anthony, 2013).

Another limitation is that our search strategy did not capture studies conducted outside of the United States or studies that were not peer reviewed (i.e., dissertations). It is possible that findings from such studies systematically differ from published studies, particularly because studies with large, visually apparent effects may be easier to publish than studies with smaller effects or more ambiguous data (Gage, Cook, & Reichow, 2017). If these trends hold in the literature, our findings might overstate the effectiveness of social stories. Future research should invest in searching for and reviewing unpublished studies to mitigate the risk of publication bias.

Future Research. Findings from this review highlight the need for powerful behavioral interventions to impact persistent challenging behaviors. Thus, researchers should examine the use of more intensive and multi-component interventions with young children in addition to the impact of the number of intervention sessions, or dosage, on child outcomes (Fey, Yoder, Warren, & Bredin-Oja, 2013). Although the method of determining adequate dosage varies across the literature (i.e., number of times or frequency of dose) (Parker-McGowan et al., 2014), it is widely understood that interventions are less efficacious when participants do not receive an adequate number of intervention sessions.

It also is important to evaluate how social story interventions conducted in early childhood settings align with contemporary standards for SCDs, as well as evaluate their effectiveness using robust effect size estimates. Although reviews have been conducted on social story interventions with older children (Karkhaneh et al., 2010; Rhodes, 2014; Sansosti et al., 2004), only three research teams (Qi et al., 2015; Wong et al., 2014; Zimmerman & Ledford, 2017) systematically reviewed how these studies adhere to the WWC standards for single case design. A few reviews reported non-overlap measures to quantify the effectiveness of social stories as a primary intervention (Kokina & Kern, 2010; Qi et al., 2015; Reynhout & Carter, 2006, 2011; Test et al., 2011), but none used parametric effect sizes or meta-analysis methods to synthesize findings. In future research, we encourage researchers to adopt parametric effect size measures, such as log response ratios, that are both interpretable and suitable for the dependent variables used in the literature. Similarly, future research should use meta-analytic models, such as multi-level random effects models, that not only summarize average effects, but also describe variability in effectiveness and examine factors that may explain this variation. Finally, future research should examine the consistency and accuracy with which a social story intervention is

implemented for young children. This would involve evaluating modifications made to social story interventions and how such modifications impact challenging behavior. Finally, as discussed earlier, it is important to note that social stories were developed to support children with ASD. Future research should investigate what groups of children and behaviors are best suited for social story interventions.

Implications for Practice. Many questions still remain unanswered as to the efficacy of social stories. Since 1993, social stories have been widely used in a variety of settings to support the social and emotional development of young children, yet moderate empirical evidence exists to support this intervention. Educators, families, and related service personnel have provided anecdotal evidence for the effectiveness of social stories with young children, but the analyses presented in this manuscript demonstrate that the effectiveness of social stories is variable. Given that studies demonstrated variability in rigor and effectiveness when social stories were implemented in isolation or as a packaged intervention, this practice should not be considered an evidence-based strategy for young children.

Social stories can be implemented in combination with other developmentally appropriate strategies, for researchers have shown that many preventive strategies can be implemented within a tiered model of support and result in decreased levels of challenging behavior (c.f., Covington-Smith, Lewis, & Stormont, 2011; Hemmeter, Snyder, Fox, & Algina, 2016). The implementation of promotion and prevention strategies highlight the need for professional development that focuses on helping teachers and support staff acquire additional skills to address challenging behavior. Employing these practices can support young children's social emotional development, as well as their overall learning and development.

Conclusion

The purpose of this systematic review was to critically evaluate the impact of social stories on young children with challenging behaviors. Results indicate variability in rigor and effectiveness of the use of social stories as an isolated intervention, as well as in combination with other intervention approaches. Thus, social stories cannot be considered an evidence-based practice for young children. Although several reviews recommend that practitioners do not use social stories as an intervention (Qi et al., 2015; Zimmerman & Ledford, 2017), while other reviews indicate that social story interventions are mildly effective (Reynhout & Carter, 2011; Rhodes, 2014; Test et al., 2011), the findings from this review highlight the need for additional research to improve our understanding of the efficacy of social story interventions in isolation and in combination with other interventions.

Supplementary Materials

Supplementary materials for this article are available on the journal's website and on the Open Science Framework at <https://osf.io/7uvr3/>. The raw data and R scripts for replicating all reported analyses are available at <https://osf.io/7uvr3/>.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by a leadership grant from the Office of Special Education Programs (Project FOCAL, H325D070061) and by Grant R305D160002 from the Institute of Educational Sciences, U.S. Department of Education. The opinions expressed are those of the author and do not represent the views of the Institute or the U.S. Department of Education.

Acknowledgements

The authors would like to acknowledge Moon Chung and Danielle Pizzella for their assistance with reliability assessment and data collection.

References

- Benish, T. M., & Bramlett, R. K. (2011). Using Social Stories™ to decrease aggression and increase positive peer interactions in normally developing pre-school children. *Educational Psychology in Practice*, 27, 1-17.
- Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I² is not an absolute measure of heterogeneity. *Research Synthesis Methods*, 8(1), 5–18. <https://doi.org/10.1002/jrsm.1230>
- Burgess, K., Chien, N., Morrissey, T., & Swenson, K. (2014). *Trends in the use of early care and education, 1995–2011: Descriptive analysis of child care arrangements from national survey data*. Washington, DC: U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation.
- Burke, R. V., Kuhn, B. R., & Peterson, J. L. (2004). Brief report: A “storybook” ending to children’s bedtime problems: The use of a rewarding Social story™ to reduce bedtime resistance and frequent night waking. *Journal of Pediatric Psychology*, 29, 389-396.
- Chan, J. M., & O'Reilly, M. F. (2008). A Social Stories™ intervention package for students with autism in inclusive classroom settings. *Journal of Applied Behavior Analysis*, 41, 405-409.
- Covington-Smith, S., Lewis, T. J., & Stormont, M. (2011). The effectiveness of two universal behavioral supports for children with externalizing behavior in Head Start classrooms. *Journal of Positive Behavior Interventions*, 13(3), 133-143.

- Crozier, S., & Tincani, M. J. (2007). Effects of social stories on prosocial behavior of preschool children with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 37, 1803-1814. doi: 10.1007/s10803-006-0315-7
- Darling-Churchill, K. E., & Lippman, L. (2016). Early childhood social and emotional development: Advancing the field of measurement. *Journal of Applied Developmental Psychology* 45, 1-7.
- Dodge, K. A., Bierman, K. L., Coie, J. D., Greenberg, M. T., Lochman, J. E., McMahon, R. J., Pinderhughes, E. E. (2014). Impact of early intervention on psychopathology, crime, and well-being at age 25. *American Journal of Psychiatry*, 172(1), 59-70.
- Fey, M. E., Yoder, P. J., Warren, S. F., & Bredin-Oja, S. L. (2013). Is more better? Milieu communication teaching in toddlers with intellectual disabilities. *Journal of Speech, Language, and Hearing Research*, 56(2), 679-693.
- Gage, N. A., Cook, B. G., & Reichow, B. (2017). Publication bias in special education meta-analyses. *Exceptional Children*, 83(4), 428–445. <https://doi.org/10.1177/0014402917691016>
- Gilliam, W. S. (2005). *Prekindergartners left behind: Expulsion rates in the state prekindergarten programs*. New Haven, CT: Yale University Child Study Center.
- Gray, C., & Garand, J. (1993). Social stories: Improving responses of students with autism with accurate social information. *Focus on Autistic Behavior*, 8(1), 1-10.
- Hedges, L. V, Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1, 39–65. <https://doi.org/10.1002/jrsm.5>
- Hemmeter, M. L., & Conroy, M. A. (2018). Advancement of evidence-based programs for young children with social and emotional learning difficulties. *School Mental Health*, 10,

199-201.

- Hemmeter, M. L., Ostrosky, M. M., & Fox, L. (2006). Social and emotional foundations for early learning: A conceptual model for intervention. *School Psychology Review, 35*, 583-601.
- Hsu, N., Hammond, H., & Ingalls, L. (2012). The effectiveness of culturally based social stories to increase appropriate behaviors of children with developmental delays. *International Journal of Special Education, 27*(1), 104-116.
- Huesmann, L. R., Dubow, E. R., & Boxer, P. (2009). Continuity of aggression from childhood to early adulthood as a predictor of life outcomes: Implications for the adolescent-limited and life-course-persistent models. *Aggressive Behavior, 35*(1), 136-149.
- Ivey, M. L., Heflin, J., & Alberto, P. (2004). The use of social stories to promote independent behaviors in novel events for children with PDD-NOS. *Focus on Autism and Other Developmental Disabilities, 19*, 164–176.
- Karkhaneh, M., Clark, B., Ospina, M. B., Seida, J. C., Smith, V., & Hartling, L. (2010). Social stories to improve social skills in children with autism spectrum disorder: A systematic review. *Autism, 14*(6), 641-662.
- Kassardjian, A., Leaf, J. B., Ravid, D., Leaf, J. A., Alcalay, A., Dale, S., . . . Oppenheim-Leaf, M. L. (2014). Comparing the teaching interaction procedure to social stories: A replication study. *Journal of Autism and Developmental Disorders, 44*, 2329-2340.
- Kokina, A., & Kern, L. (2010). Social story™ intervention for students with autism spectrum disorders: A meta-analysis. *Journal of Autism and Developmental Disorders, 40*, 812-826.
- Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34*(1), 26-38. <http://dx.doi.org/10.1177/0741932512452794>.

- Kuoch, H., & Mirenda, P. (2003). Social story™ interventions for young children with autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities, 18*(4), 219-227.
- Laprime, A. P., & Dittrich, G. A. (2014). An evaluation of a treatment package consisting of discrimination training and differential reinforcement with response cost and a Social story on vocal stereotypy for a preschooler with autism in a preschool classroom. *Education and Treatment of Children, 37*(3), 407-430.
- Leaf, J. B., Oppenheim-Leaf, M. L., Call, N. A., Sheldon, J. B., & Sherman, J. A. (2012). Comparing the teaching interaction procedure to social stories for people with autism. *Journal of Applied Behavior Analysis, 45*, 281-298.
- Leaf, J. B., Oppenheim-Leaf, M. L., Leaf, R. B., Taubman, M., McEachin, J., Parker, T., . . . Mountjoy, T. (2015). What is the proof? A methodological review of studies that have utilized social stories. *Education and Training in Autism and Developmental Disabilities, 50*, 127-141.
- Lorimer, P. A., Simpson, R. L., Myles, B. S., & Ganz, J. B. (2002). The use of social stories as a preventive behavioral intervention in a home setting with a child with autism. *Journal of Positive Behavior Interventions, 4*(1), 53-60.
- Losel, F., & Bender, D. (2012). Child social skills training in the prevention of antisocial development and crime. In D. Farrington, & B. C. Welsh (Eds.), *The Oxford handbook of crime prevention* (pp. 102–129). Oxford, UK: Oxford University Press.
- Mayton, M. R., Menendez, A. L., Wheeler, J. J., Carter, S. L., & Chitiyo, M. (2013). An analysis of social stories research using an evidence-based practice model. *Journal of Research in Special Education Needs, 13*, 208-217.
- Moeyaert, M., Maggin, D. M., & Verkuilen, J. (2016). Reliability, validity, and usability of data extraction programs for single-case research designs. *Behavior Modification, 40*(6), 874-900.

<https://doi.org/10.1177/0145445516645763>

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *PLoS Med* 6(7), e1000097.

<https://doi.org/10.1371/journal.pmed.1000097>

Ostrosky, M. M., & Sandall, S. R. (Eds.) (2013). *Addressing young children's challenging behaviors* (Young Exceptional Children Monograph Series No. 15). Los Angeles, CA: The Division for Early Childhood of the Council for Exceptional Children.

Parker-McGowan, Q., Chen, M., Reichle, J., Pandit, S., Johnson, L., & Kreibich, S. (2014).

Describing treatment intensity in milieu teaching interventions for children with developmental disabilities: A review. *Language, Speech, and Hearing Services in Schools*, 45, 351-364.

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy*, 42(2), 284–299.

<https://doi.org/10.1016/j.beth.2010.08.006>

Powell, D., Fixsen, D., Dunlap, G., Smith, B., & Fox, L. (2007). A synthesis of knowledge relevant to pathways of service delivery for young children with or at risk of challenging behavior. *Journal of Early Intervention*, 29(2), 81-106.

Pustejovsky, J. E. (2017). clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections. Retrieved from <https://cran.r-project.org/package=clubSandwich>

Pustejovsky, J. E. (forthcoming). Procedural sensitivities of effect sizes for single-case designs with directly observed behavioral outcome measures. *Psychological Methods*.

<https://doi.org/10.1037/met0000179>

Pustejovsky, J. E. (2018b). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. *Journal of School Psychology*, 68, 99–112.

<https://doi.org/10.1016/j.jsp.2018.02.003>

Pustejovsky, J. E., & Ferron, J. M. (2017). *Research synthesis and meta-analysis of single-case designs. Handbook of Special Education: Second Edition.*

<https://doi.org/10.4324/9781315517698>

Pustejovsky, J. E., & Swan, D. M. (2018). SingleCaseES: A calculator for single-case effect sizes. Retrieved from <https://github.com/jepusto/SingleCaseES>

Qi, C. H., Barton, E. E., Collier, M., Lin, Y. L., & Montoya, C. (2015). A systematic review of effects of social stories interventions for individuals with autism spectrum disorder. *Focus on Autism and Other Developmental Disabilities*, 1-10.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.r-project.org/>

Raver, S. A., Bobzien, J., Richels, C., Hester, P., & Anthony, N. (2013). Using dyad-specific social stories to increase communicative and social skills of preschoolers with hearing loss in self-contained and inclusive settings. *International Journal of Inclusive Education* 18(1), 18-35.

Reynhout, G., & Carter, M. (2006). Social Stories™ for children with disabilities. *Journal of Autism and Developmental Disorders*, 36, 445-469. doi: 10.1007/s10803-006-0086-1

Reynhout, G., & Carter, M. (2011). Evaluation of the efficacy of Social Stories™ using three single subject metrics. *Research in Autism Spectrum Disorders*, 5, 885-900.

Rhodes, C. (2014). Do social stories help to decrease disruptive behavior in children with autism spectrum disorders? A review of published literature. *Journal of Intellectual Disabilities*, 18(1), 35-50. doi: 10.1177/1744629514521066

Rohatgi, A. (2018). WebPlotDigitizer (Version 4.1). Retrieved from <https://automeris.io/WebPlotDigitizer>

- Sansosti, F. J., Powell-Smith, K. A., & Kincaid, D. (2004). A research synthesis of Social Story™ interventions for children with autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities*, 19(4), 194-204.
- Schneider, N., & Goldstein, H. (2010). Using social stories and visual schedules to improve socially appropriate behaviors in children with autism. *Journal of Positive Behavior Interventions*, 12(3), 149-160.
- Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education*, 8(2), 24–43. <https://doi.org/10.1177/074193258700800206>
- Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention*, 2(3), 188–196. <https://doi.org/10.1080/17489530802581603>
- Shonkoff, J. P. (2016). Capitalizing on advances in science to reduce the health consequences of early childhood adversity. *JAMA Pediatrics*, 170, 1003-1007.
- Soenksen, D., & Alper, S. (2006). Teaching a young child to appropriately gain attention of peers using a Social story™ intervention. *Focus on Autism and Other Developmental Disabilities*, 21(1), 36-44.
- Tarlow, K. R. (2017). An improved rank correlation effect size statistic for single-case designs: Baseline corrected Tau. *Behavior Modification*, 41(4), 427–467.
<https://doi.org/10.1177/0145445516676750>
- Test, D. W., Richter, S., Knight, V., & Spooner, F. (2011). A comprehensive review and meta-analysis of the Social Stories™ literature. *Focus on Autism and Other Developmental Disabilities*, 26(1), 49-62. doi: 10.1177/1088357609351573

- Tipton, E. & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics, 40*(6), 604-634. <https://doi.org/10.3102/1076998615606099>
- Tremblay, R. E. (2010). Developmental origins of disruptive behavior problems: The “original sin” hypothesis, epigenetics, and their consequences for prevention. *Journal of Child Psychology and Psychiatry, 51*, 341–367.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48.
- Wong, C., Odom, S. L., Hume, K. Cox, A. W., Fettig, A., Kucharczyk, S., ... Schultz, T. R. (2014). *Evidence-based practices for children, youth, and young adults with autism spectrum disorder*. Chapel Hill: The University of North Carolina, Frank Porter Graham Child Development Institute, Autism Evidence-Based Practice Review Group.
- Wright, L. A., & McCathren, R. B. (2012). Utilizing social stories to increase prosocial problem behavior in young children with autism. *Child Development Research, 1*-13.
- Zimmerman, K. N., & Ledford, J. R. (2017). Beyond ASD: Evidence for the effectiveness of social narratives. *Journal of Early Intervention, 1*-19.

Figure 1

PRISMA Flow Diagram: Included & Excluded Studies for Literature Review

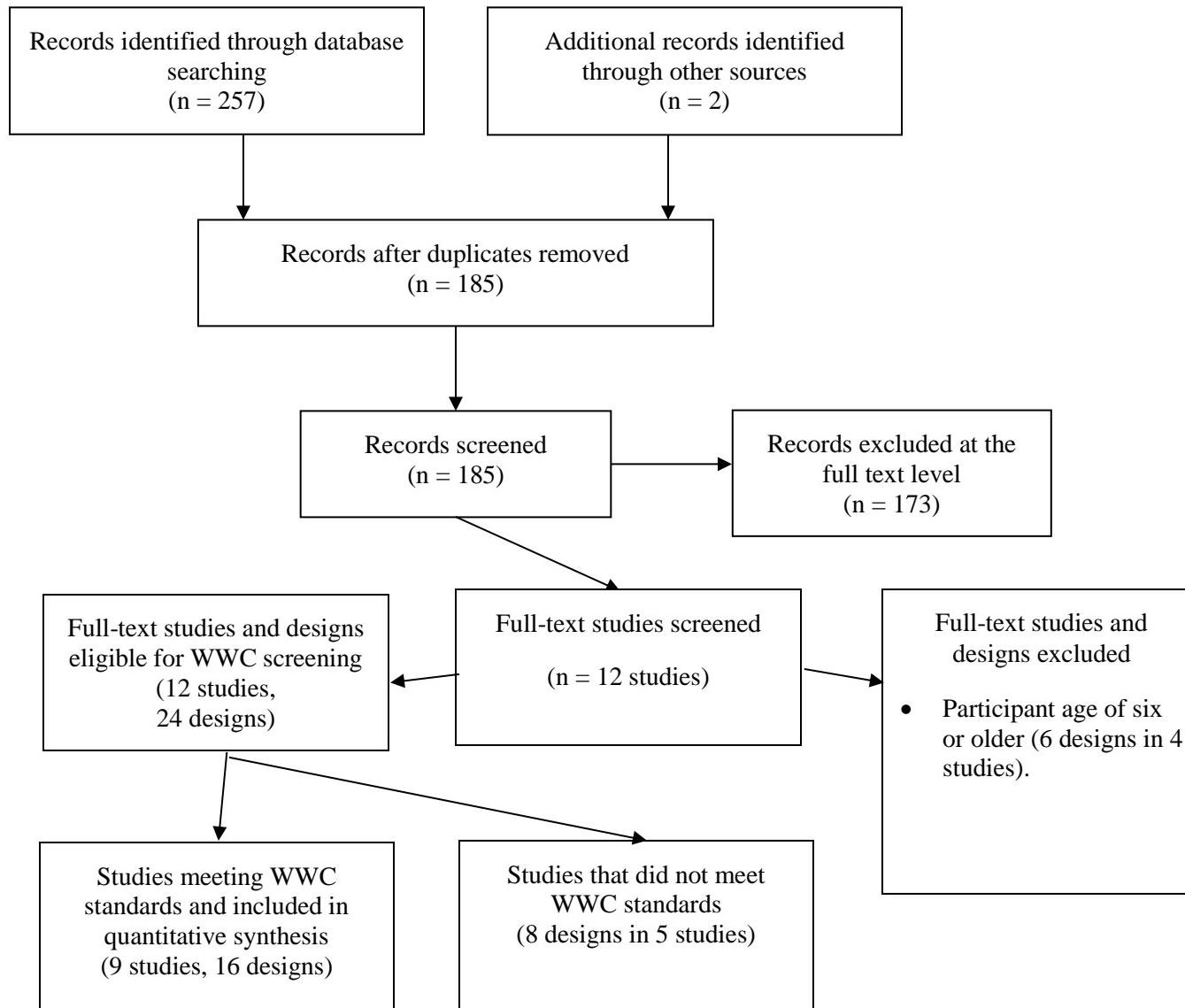


Table 1
Study Summary, Overall Design Rating, & Strength of Evidence of Experimental Control

Author(s)	Participants			Target behavior (DV)	IV	Design	Overall Design	Evidence Rating
	Age	Gender	Diagnosis					
Benish & Bramlett (2011)	4 YO* = 3	F = 1 M = 2	None	Avoidance, refusal, physical & verbal aggression, name calling, hitting, pushing, not sharing	Social Stories™	MBP	Does not meet	No evidence
Burke, Kuhn, & Peterson (2004)	7 YO = 2 5 YO = 1 2 YO = 1	F = 2 M = 2	None	Tantrums, hitting, kicking, destruction of property, frequent night waking, difficulty initiating & maintaining sleep, fighting, arguing, crying, screaming, waking, entering parents' bed during the night, refusing to go to bed	Social stories, tangible rewards	ABAB MBP	Does not meet Meets w/reservations	No evidence Moderate
Chan & O'Reilly (2008)	5 YO = 1	M = 1	ASD	appropriate hand raising, inappropriate vocalizations, appropriate social initiations	Social Stories™, role play, answering questions	Multi-probe across behaviors	Does not meet	No evidence
Crozier & Tincani (2007)	5 YO = 1 3 YO = 2	M = 3	ASD	Sitting appropriately during circle time, talking with peers during snack time, playing appropriately with peers	Social Stories™, verbal prompts	ABAB ABCAC BC ABAB ABAB	Meets w/reservations Does not meet Meets standards Meets standards	Moderate No evidence Moderate No evidence
Hsu, Hammond, & Ingalls (2012)	6 YO = 1	M = 3	ASD SLI DD	Sitting at seat, following directions, raising one's hand	Social Stories™	MBP	Meets w/reservations	No evidence

	4 YO = 1 3 YO = 1							
Ivey et al. (2004)	5 YO = M = 2 1 5 YO = 1	ASD	Using necessary materials, following directions, following the rules of a game, make requests, use target vocabulary, remain on task	Social Stories™	ABAB ABAB	Meets w/reservations Meets w/reservations	No evidence No evidence	
Kuoch & Mirenda (2003)	5 YO = M = 1 2 3 YO = 1	ASD	Aggression, crying, yelling, screaming, squealing, throwing up food, placing hands on genitals	Social Stories™	ABA ABA	Does not meet Does not meet	No evidence No evidence	
Leaf et al. (2012)	5 YO = M = 3 3	ASD	Losing/wining graciously, sportsmanship, giving compliments, cheering up a friend, appropriate greetings, changing the conversation,	Social Stories™ Teacher Interaction Procedure	MBB (Fig. 2-4)	Meets w/reservations	No evidence	
Lorimer, Simpson, Myles, & Ganz (2002)	5 YO=1 M = 1	ASD	Precursors to tantrum behavior (e.g., screaming, hitting, kicking, & throwing objects)	Social Stories™	ABAB ABAB	Meets w/reservations Meets w/reservations	Moderate Moderate	

Schneider & Goldstein (2010)	10 YO = = 1 6 YO = = 1 5 YO = 1	M = 3	ASD	Wandering around the classroom, working at a computer after the bell rang, standing next to a student & leaning over to look at the computer screen while standing in line, touching or leaning over to look at a computer, speaking without raising hand or not waiting to be called on, rolling or lying on the ground, leaving the group situation without being asked by the teacher, looking away from the teacher & not following directions	Social Stories™	MBP	Meets standards	Moderate
Soenksen & Alper (2006)	5 YO = 1	M = 1	Hyperlexia	Verbally saying a peer's name, looking at a peer's face	Social Stories™	MBA	Meets w/reservations	Moderate
Wright & McCathren (2012)	4 YO = 2 5 YO = 2	M = 4	ASD	Positive peer interaction; response to peer initiation	Social Stories™	MBP	Meets standards	No evidence

Note: YO = Years Old; M = Male; F = Female; ASD = Autism Spectrum Disorder; SLI = Specific Language Impairment; DD = Developmental Delay; ADHD = Attention Deficit Hyperactivity Disorder; DV = Dependent variable; IV = Independent variable; MBP = Multiple-baseline across participants; MBB = Multiple baseline across behaviors; MBA = Multiple baseline across activities

Table 2

Meta-analysis and Moderator Analysis of Log Response Ratio-increasing Effect Size Estimates

Predictor	Studies	Effect sizes	Average LRRi Estimate (SE)	Average LRRi 95% CI	Study-level SD	Case-level SD	Test of between-group differences
Summary meta-analysis							
Overall average effect	9	39	1.01 (0.15)	[0.65, 1.37]	0.00	1.05	
Challenging behavior	4	10	1.22 (0.44)	[-0.20, 2.63]	0.65	0.71	
Prosocial behavior	7	29	0.94 (0.19)	[0.44, 1.43]	0.00	1.10	
Participant age in years							
Five or younger	9	34	1.01 (0.10)	[0.75, 1.27]	0.00	1.07	$F(1, 2.4) = 0.00, p = .95$
Six or older	3	5	0.97 (0.64)	[-2.07, 4.02]			
Participant diagnosis							
Diagnosed disability	7	90	0.89 (0.16)	[0.46, 1.31]	0.00	1.05	$F(1, 1.76) = 0.59, p = .53$
No diagnosed disability	2	9	1.36 (0.60)	[-6.23, 8.95]			
WWC design rating							
Meets standards	3	13	0.94 (0.30)	[-0.61, 2.49]	0.00	0.98	$F(2, 3.3) = 0.77, p = .53$
Meets standards with reservations	7	26	1.03 (0.19)	[0.50, 1.57]			
Does not meet standards	5	15	1.40 (0.23)	[0.42, 2.38]			
Interventionist & primary data collector							
Different	3	13	0.81 (0.40)	[-1.30, 2.93]	0.00	1.05	$F(1, 3.2) = 0.41, p = .57$
Same	6	26	1.09 (0.18)	[0.58, 1.60]			

Notes: SE = standard error. CI = confidence interval. SD = standard deviation.